Original Investigation

Transfer Learning From Convolutional Neural Networks for Computer-Aided Diagnosis: A Comparison of Digital Breast Tomosynthesis and Full-Field Digital Mammography

Kayla Mendel, Hui Li, Deepa Sheth, Maryellen Giger

Abbreviations

ARD architecture distortion
AUC area under the ROC curve
CADe computer-aided detection
CADx computer-aided diagnosis
CC cradiocaudal
CNN convolutional neural network
DBT digital breast tomosynthesis
FFDM full-field digital mammography
HIPAA health insurance portability and accountability act
MLO mediolateral oblique
ROC receiver operating characteristic
ROI region of interest
SVM support vector machine

Rationale and Objectives: With the growing adoption of digital breast tomosynthesis (DBT) in breast cancer screening, we compare the performance of deep learning computer-aided diagnosis on DBT images to that of conventional full-field digital mammography (FFDM).

Materials and Methods: In this study, we retrospectively collected FFDM and DBT images of 78 biopsy-proven lesions from 76 patients. A region of interest was selected for each lesion on FFDM, synthesized 2D, and DBT key slice images. Features were extracted from each lesion using a pretrained convolutional neural network (CNN) and served as input to a support vector machine classifier trained in the task of predicting likelihood of malignancy.

Results: From receiver operating characteristic (ROC) analysis of all 78 lesions, the synthesized 2D image performed best in both the cradiocaudal view (area under the ROC curve [AUC] = 0.81, SE = 0.05) and mediolateral oblique view (AUC = 0.88, SE = 0.04) in the task of lesion characterization. When cradiocaudal and mediolateral oblique data of each lesion were merged through soft voting, DBT key slice image performed best (AUC = 0.89, SE = 0.04). When only masses and architectural distortions (ARDs) were considered, DBT performed significantly better than FFDM ($p = 0.024$).

Conclusion: DBT performed significantly better than FFDM in the merged view classification of mass and ARD lesions. The increased performance suggests that the information extracted by the CNN from DBT images may be more relevant to lesion malignancy status than the information extracted from FFDM images. Therefore, this study provides supporting evidence for the efficacy of computer-aided diagnosis on DBT in the evaluation of mass and ARD lesions.

Keywords: Digital breast tomosynthesis; Mammography; Deep learning; Convolutional neural networks; Computer-aided diagnosis.

Acad Radiol 2018; ■:1–9

From the The University of Chicago, 5801 S Ellis Ave, Chicago, Illinois. Received May 16, 2018; revised June 13, 2018; accepted June 22, 2018. **Address correspondence to:** K.M. e-mail: kmendel@uchicago.edu

https://doi.org/10.1016/j.acra.2018.06.019

INTRODUCTION

Digital breast tomosynthesis (DBT) has emerged as a promising modality to improve screening sensitivity and accuracy. DBT produces pseudo-3D images by rotating an x-ray source in a partial arc around the breast

while acquiring projection images. A growing number of studies have shown that tomosynthesis significantly reduces screening recall rates and increases cancer detection rates (1–4). By providing volume data as opposed to single projection images, DBT gives a clearer visualization of regions of interest by minimizing overlaying tissue compared to 2D full field digital mammography. Therefore, DBT is expected to be particularly useful for women with dense breasts for whom overlaying parenchymal tissue may obscure breast lesions (5). However, human observer studies are inherently qualitative and subjective interpretations. The objectivity of computer vision methods may therefore help inform imaging strategies across breast radiology.

The growing adoption of DBT in screening protocols makes the prospect of computer-aided diagnosis (CADx) on DBT images clinically impactful. Therefore, it is informative to compare performance on DBT to that on full-field digital mammography (FFDM). Several groups have studied computer-aided detection) of lesions using DBT images with conventional radiomic methods, yielding promising results (6–8). These conventional methods are being superseded in some applications by emerging artificial intelligence approaches such as deep learning.

Deep learning is a machine learning method which is rapidly growing in usage in the image processing field. Deep convolutional neural networks (CNNs) have seen the most widespread use in object detection and image classification tasks. These methods involve computing high dimensional, unintuitive features from large databases. This contrasts with previous CADx and computer-aided detection research which compute relatively small numbers of handcrafted intuitive features as CNNs can extract features through convolutional, pooling, and connected layers (9,10).

Deep learning is now being used in medical imaging classification tasks (11,12). Compared to natural object sets such as ImageNet (13), annotated medical datasets are limited in size. To handle small databases, approaches for medical classification tasks typically involve transfer learning through the application of a pretrained CNN. The pretrained CNN is typically intended for multiclass object classification on a database such as ImageNet, as illustrated in Figure 1 (14). Essentially, pretrained neural networks act as feature extractors for image sets in different domains. Different domains typically have different population characteristics and different classification categories, thus necessitating a classifier such as support vector machine (SVM) (15,16). Transfer learning has been applied in DBT lesion detection tasks, with applications on detecting both masses and calcifications (17,18). Transfer learning has also been applied to lesion characterization with DBT, however comparison across image types was not performed (19).

In order to compare the efficacy of transfer-learning based CADx on DBT and FFDM, transfer parameters were used to build classification models for each image type. Evaluation of the performance of deep learning features on FFDM and DBT images may provide further support in the utilization of extending deep learning-based CADx to DBT applications. This may improve the precision and accuracy of characterizing breast lesions. The aim of this study is to provide an objective comparison between the diagnostic performance of FFDM, synthesized 2D image, and DBT key slice in the tomosynthesis cine loop through CADx in differentiating malignant from benign breast lesions. This type of comparison is innovative as while it is common to compare performance over different algorithms, comparison of performance across different image types is relatively unexplored.

MATERIALS AND METHODS

Database description

A retrospective review was performed on all patients who had undergone both FFDM and DBT resulting in a mammographically-detected lesion which was ultimately biopsied for final surgical pathology. FFDM and DBT imaging were performed on a Hologic Selenia Dimensions unit (Marlborough, Massachusetts). All aspects of the diagnostic workup were performed at the University of Chicago Medicine, and images were retrospectively collected under Health Insurance Portability and

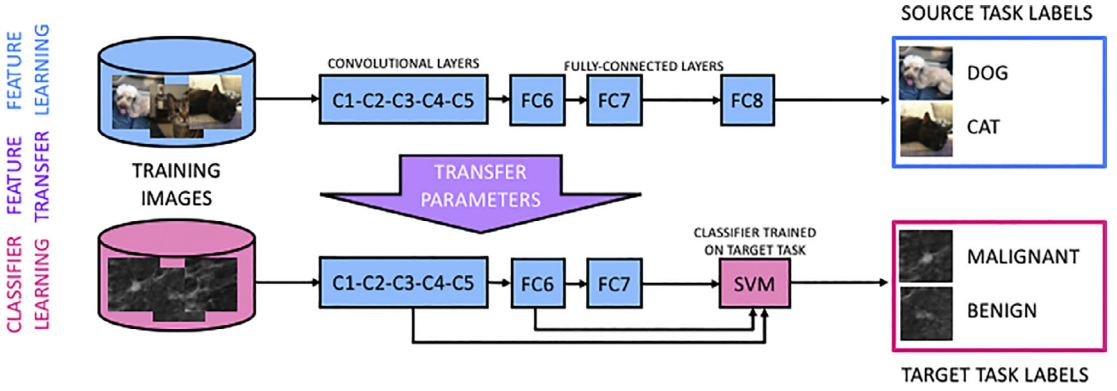

Figure 1. Illustration of the general deep learning approach of transfer learning through feature extraction. Parameters are transferred from a pretrained neural network. Features are then extracted from the various layers on images from a separate domain, such as medical imaging.

TABLE 1. Summary of Patient Ages, Lesion Types, and Lesion Molecular Subtypes

	Frequency (%)	
	Malignant	Benign
Age		
≤39	–	1 (2.1)
40–49	6 (20.0)	20 (41.7)
50–59	7 (23.3)	18 (37.5)
60–69	12 (40.0)	9 (18.8)
≥70	5 (16.7)	–
Average age (SD)	59.6 (10.3)	51.5 (8.6)
Lesion type		
Mass	10 (33.3)	23 (47.9)
Architectural distortion (ARD)	9 (30.0)	6 (12.5)
Calcifications	11 (36.7)	19 (39.6)
Molecular subtype		
DCIS	14	
IDC	12	
ILC	3	
Invasive mammary	1	
Papillary carcinoma	1	
Atypical ductal hyperplasia (ADH)		7
Complex sclerosing lesion		3
Fibroadenoma (FA)		9
Fibrocystic change		7
Normal breast parenchyma		5
Cyst		1
Apocrine metaplasia		4
Stromal fibrosis		3
Intraductal papilloma		5
Sclerosing adenosis		3
Usual ductal hyperplasia (UDH)		1
Total	30	48

Accountability Act approved and institutional review board approved protocols. A total of 76 patients with 78 lesions were included in this study, with exams ranging in date from August 2015 to June 2017. The average age of included patients was 54.7 years (standard deviation = 10.1 years). Of the 78 lesions, 30 lesions were biopsy proven to be malignant and 48 lesions were biopsy proven to be either high risk or benign. A summary of patient and lesion characteristics is included in Table 1. Each lesion was identified in the CC and MLO view on the (1) FFDM image, (2) synthesized 2D image, and (3) DBT key slice in the tomosynthesis cine loop. A fellowship-trained breast imager manually identified the key slice of each lesion from the tomosynthesis cine loop. For mass lesions, the key slice was defined as the slice nearest to the center of the lesion with the largest lesion diameter and/or when the lesion was best in focus. For architectural distortion lesions, the key slice was defined as that in which the largest number of spiculations were seen and/or when the lesion was best in focus. For calcification lesions, the key slice was defined as that in which the greatest number of calcifications were in focus. We acknowledge that manual selection of a key slice may introduce bias in analysis, particularly if the mass lesion is not circumscribed or the calcifications are not along the plane of image acquisition. In future work, methods of evaluating the full lesion volume will be explored. These may have the potential to further improve classification performance beyond that observed in this study.

In terms of lesion categorization, the high risk lesions and the benign lesions were grouped into the "benign" category. All high risk lesion patients either ultimately underwent surgical excision or had at least two years of imaging follow-up. No patients were upgraded to malignancy in this high risk lesion category.

Feature extraction and reduction

The VGG19 deep convolutional neural network, which consists of 19 weight layers, was used to extract features in this study (20). VGG19 was pretrained on over one million images from the ImageNet dataset which consists of natural objects used for multiclass object classification (10,13). Learned weights obtained during pretraining were applied to the network in this study, and features were extracted from various layers of the network. These features were used in the research task of classifying breast lesions as malignant or benign. Note that due to the small database size, the network was not trained or fine-tuned in order to avoid overfitting. Instead, images were fed through the existing architecture, and quantitative features were extracted from various layers (21).

A fellowship-trained breast radiologist identified each lesion on all three modalities: FFDM, DBT synthesized image, and DBT key slice. A square region of interest (ROI) measuring 512×512 pixels was manually placed to fully cover the lesion on both the CC and MLO views for each image type. ROIs were then bicubically interpolated to a size of 224×224 pixels to conform to the size of training images used in the initial training of VGG19. Examples of malignant and benign ROIs from each image type is shown in Figure 2.

Features were extracted from each max pooling layer of the VGG19 convolutional network for each of these modalities, and features from each maxpool layer were fed through a meanpool layer to reduce the number of features. Following initial feature extraction, feature dimension reduction was further conducted by eliminating features with zero variance over all lesions considered in this study.

Feature Selection and Classification

Following feature extraction and reduction, leave-one-out stepwise feature selection was performed to identify a nonredundant set of informative features (22). To identify such a feature set, stepwise feature selection was performed in a leave-one-out manner over the training data, with one training case left out each round. Each round, stepwise feature selection was performed by iteratively adding and removing

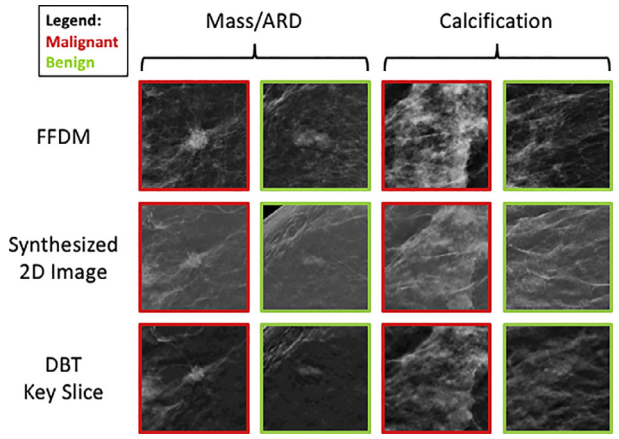

Figure 2. Examples of malignant and benign regions of interest selected to use for classification of two masses and two calcifications. Regions of interest for the four example lesions are shown in each of the image types explored in this paper.

features from the classification feature set, using the *p* value of the F-statistic as a metric to measure significance in improvement of the model. The null hypothesis is that a candidate feature would have a zero coefficient in the multilinear model, and if there exists sufficient evidence to reject the null hypothesis, then the candidate feature is added to the model. Conversely, if there exists insufficient evidence to reject the null hypothesis, then the candidate feature is removed from the model. This iterative algorithm continues until no single step improves the model. Stepwise feature selection is described in greater detail elsewhere (22).

After repeating the stepwise feature selection algorithm for reach left out training case, the cumulative frequency of the selection of individual features is considered. The most frequently selected features over the leave-one-out iterations were selected for use in the classification model. The motivation behind this iterative method of feature selection is to keep the number of features included in models constant when comparing across image types. Stepwise feature selection on its own produces variable quantities of selected features, which might introduce bias into the evaluation of the performance of classifiers, as it has been shown that classification performance varies with the number of included features (23). By using the frequency of selection to identify a fixed number of features, this potential source of bias was reduced.

For combined analysis of masses and calcifications, the four most frequently selected features were used in the final classifier. For analysis of either mass/architecturural distortions or calcifications, the two most frequently-selected features were maintained. The numbers of features used in this study were selected to be near the optimal number of features for classification with SVM for the dataset size based on recommendations by Hua et al (24). The reduced feature set was used to train an SVM classifier with a linear kernel in a leave-one-out manner (25). Outputs from the SVM were used to perform receiver operating characteristic analysis and to determine the area under the receiver operating characteristic curve (AUC), which was used as a figure of merit in this study (26). The standard error of the AUC was calculated to estimate the range of values for the population. Note that analysis was performed in a leave-one-out manner as opposed to independent training and testing sets due to the small size of available data. The resulting classification performances reported in this study are therefore viewed as an overestimation of performance, as separated training and testing may yield lower performance. However, the aim of this study was to compare performance, this likely has a minimal impact on the study's conclusions.

An extension of this analysis was needed to further understand the value of complementary information provided by the two standard screening views of each breast. To this end, we investigated the performance of a merged classifier by combining signatures from the CC and MLO views of each lesion. The merged classifier was constructed through soft voting of the SVM output of classifiers trained separately on the CC and MLO view images (27). This analysis was repeated separately for each of the three imaging modalities

The visual characteristics of masses and calcifications vary, and this analysis sought to explore whether corresponding characteristics of malignancy vary as well (28). Thus, the imaging data were additionally examined in subsets based on lesion type (mass/architectural distortion or calcifications). Training and classification was repeated following the same methodology as when performed on the full dataset.

Statistical Evaluation

Statistical significance of the difference of each task's AUC from random guessing was calculated for each classifier using a statistical z test (29,30). The statistical significance of the difference between classifiers was evaluated using the p value of a univariate z-score statistical test calculated using ROCKIT software (26). Corrections for multiple comparisons were performed following the Holm–Bonferroni method (31).

RESULTS

Lesion Characterization by Single View

The AUC was determined for the classification of malignant and benign lesions for each breast imaging modality (FFDM and DBT) and for each view (CC and MLO). The resulting AUC and standard error values are presented in Table 2. For the MLO view, the performance of synthesized 2D images was higher than the performance of FFDM or DBT key slice for both calcification lesions and mass/ARD lesions. For the CC view, the

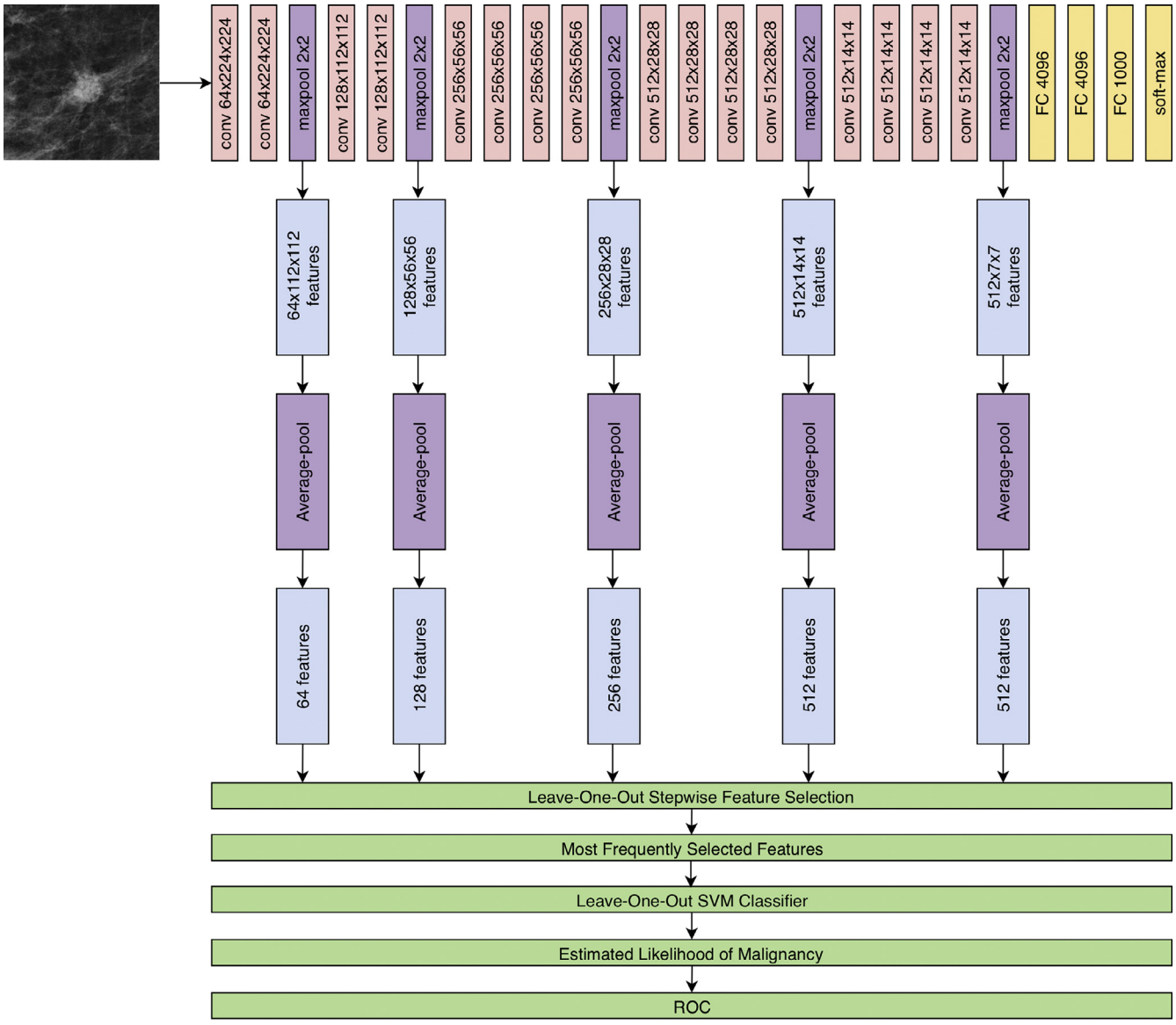

Figure 3. Structure of the VGG19 convolutional neural network, and illustration of the layers from which features were extracted and input to the SVM classifier to yield an output classification decision in this study. Features were extracted from each maxpool layer and then put through an average-pool layer to reduce feature dimensionality. Feature reduction was performed, and remaining features were input to a leave-one-out SVM classifier to produce a final classifier.

SVM, support vector machine.

TABLE 2. Summary of AUC Values Observed for Classifying Lesions as Malignant or Benign

Images analyzed		All (n = 78)	Masses/ARD (n = 48)	Calcifications (n = 30)
FFDM	**CC and MLO**	0.81 ± 0.05	0.88 ± 0.05	0.88 ± 0.06
	CC view	0.76 ± 0.05	0.90 ± 0.07	0.83 ± 0.08
	MLO view	0.76 ± 0.06	0.82 ± 0.06	0.82 ± 0.08
Synthesized 2D image	**CC and MLO**	0.86 ± 0.04	0.91 ± 0.04	0.94 ± 0.04
	CC view	0.81 ± 0.05	0.75 ± 0.08	0.88 ± 0.10
	MLO view	0.88 ± 0.04	0.87 ± 0.06	0.90 ± 0.06
DBT	**CC and MLO**	0.89 ± 0.04	0.98 ± 0.01	0.97 ± 0.03
	CC view	0.74 ± 0.05	0.79 ± 0.08	0.82 ± 0.08
	MLO view	0.83 ± 0.05	0.80 ± 0.07	0.84 ± 0.07

AUC, area under the ROC curve; CC, cradiocaudal; DBT, digital breast tomosynthesis; FFDM, full-field digital mammography; MLO, mediolateral oblique.

performance of synthesized 2D images was highest for calcification lesions, and performance of FFDM was highest for mass/ARD lesions.

Lesion Characterization by Merged Views

Lesions may be best characterized in one of the two standard views used in screening mammography (CC and MLO). Therefore, incorporation of information from both views may provide complementary information motivating this study's use of a merged classifier.

The merged classifier for DBT key slice images consistently outperformed DBT key-slice single view classifiers in each lesion subset, suggesting that the two views of DBT images provide complementary information. For FFDM and synthesized 2D images, the merged classifier did not consistently perform better than single-view classifiers. Thus, the merged classifier was not decidedly preferred on this dataset for these image types. Examples of lesions which were correctly and incorrectly classified by the various classifiers are shown in Figure 4.

Performance of the merged classifier is reported in Table 2 and illustrated in Figures 5 and 6. Performance of each the Synthesized 2D images and DBT key slices were compared to FFDM in the task of lesion characterization, using the merged-view classifiers. After correcting for multiple comparisons through the Holm–Bonferroni method, the performance of DBT key slice was significantly superior to the performance of FFDM.

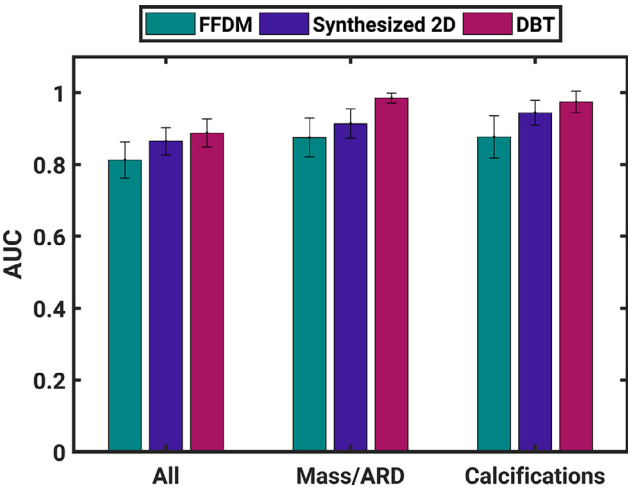

Figure 5. Classification performance of the merged-view classifier on each subset of lesions considered in this study. AUC is plotted with error bars showing one standard error.
AUC, area under the ROC curve.

DISCUSSION

In this study, we explored the potential of using pre-trained CNNs via feature extraction for the task of classifying malignant from benign breast lesions on (a) FFDM, (b) synthesized 2D images, and (c) DBT key slice images. To our knowledge, comparisons of CNN transfer learning performance across mammographic imaging modalities for lesion diagnosis has not yet been conducted. With the growing presence of DBT in breast cancer screening, it is increasingly important to understand differences between the use of these modalities in CAD algorithms such as deep learning.

The application of CNNs to classifying lesions on FFDM and DBT images as either malignant or benign was explored. Transfer learning methodology was applied, using a pre-trained CNN to extract features from FFDM and DBT ROIs. The extracted features were input to a support vector machine classifier and the diagnostic performance of resulting class probabilities was determined in terms of AUC.

When mass and ARD lesions are considered, DBT performed significantly better than FFDM in the task of classifying lesions as malignant or benign. This is in agreement with observer studies and conventional radiomics studies which also found that DBT had similar or better performance than FFDM in the task of lesion characterization (2,3,32–34).

The increased lesion conspicuity in DBT is greatly beneficial when imaging dense breast tissue because it is can be difficult to perceive suspicious lesions in extremely dense breast

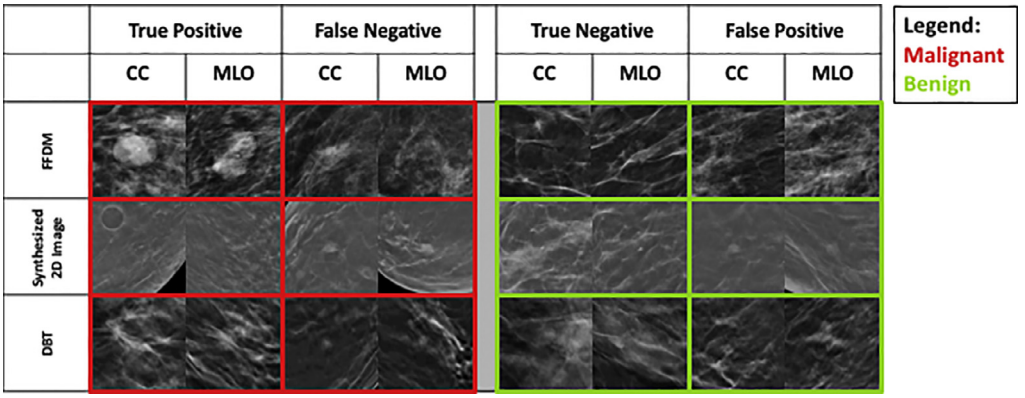

Figure 4. ROIs of lesions that were correctly or incorrectly classified by classifiers trained for each image type. The most extreme lesion (ie highest or lowest probability of malignancy) was used to select the representative lesion shown here for illustrative purposes.
ROIs, region of interests.

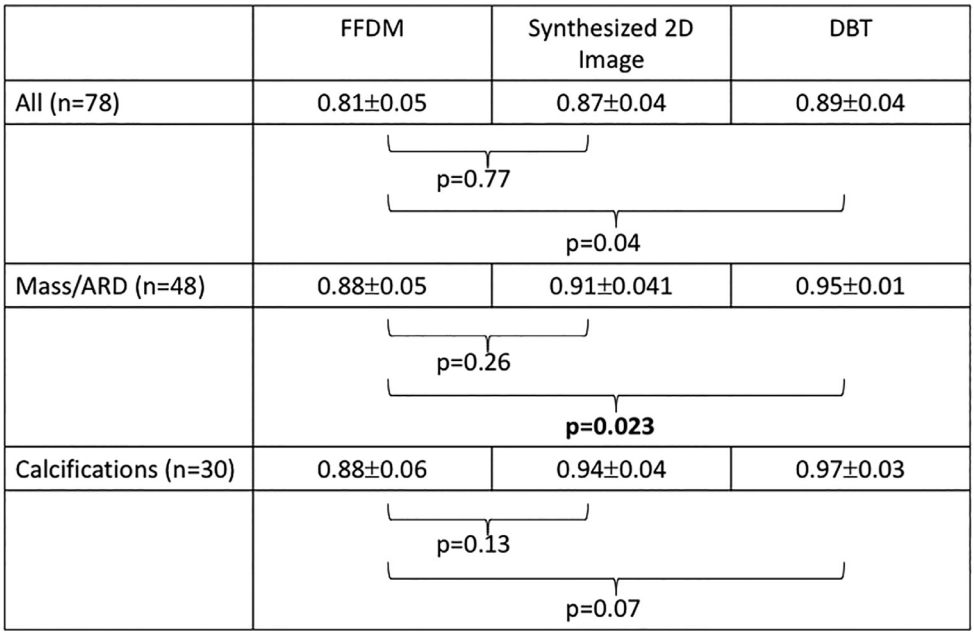

Figure 6. Significance of difference between AUC values using merged CC and MLO data for classification in the task of predicting malignancy. After corrections for multiple comparisons, a *p* value of 0.025 is significant at the $\alpha = 0.05$ significance level (31).
AUC, area under the ROC curve; CC, cradiocaudal; MLO, mediolateral oblique.

tissue. The border of masses, number of masses, and associated findings such as dilated ducts or vessels around a mass are better depicted on DBT images, especially in dense breasts (35). Because of the ability of DBT to reduce tissue superimposition, a benefit of DBT is a reduction in the recall rate in women with dense breasts. Haas et al (4) reported that the addition of DBT reduced recall rates for all breast density groups and age groups, with significant differences in recall rates for scattered heterogeneously dense and extremely dense breasts. Their study findings reiterate the belief that DBT will prove to be beneficial for patients with dense breast tissue and for those with nondense breast tissue. In our dataset, the DBT key slice yielded the highest AUC when individually classifying masses/ARD and calcifications, confirming that tomosynthesis is indeed helpful to reduce overlapping parenchymal tissue in the analysis/classification of lesions. We acknowledge that the feature dimensionality was large compared to the number of lesions included in this study. Therefore, the results reported here are treated as initial findings, and warrant further investigation on a larger dataset.

Most findings at DBT are apparent on both the CC and MLO projections, but one-view-only findings occur at DBT, and breast cancer still occasionally may be visible on only one projection. Previous studies involving DBT have estimated that 5%–9% of breast malignancies are seen only on the CC projection, whereas 1%–2% of breast malignancies are apparent only on the MLO projection (36). Moreover, 12%–15% of findings noted on both projections are more readily apparent on one view compared to the other (37). In our study, all three imaging modalities performed better when the CC and MLO views were merged, confirming that each view provides unique and synergistic information to aid in the classification of a lesion. When individually assessed, DBT synthesized 2D image performed better than all other imaging methods in both the CC and MLO view.

Resulting classification performances observed in this study were comparable to those reported by the limited number of DBT-based deep learning CADx studies. For example, while an independent data set was used, Samala et al. observed an AUC of 0.90 in the task of classifying mass lesions through an evolutionary pruning approach (19). However, this study did not compare classification performance of DBT to that of FFDM or 2D synthesized images. Therefore, while the study by Samala et al provides support for the feasibility of transfer learned deep CNNs for CADx on DBT images, the results observed in our study complement this finding by comparing performance over different image types. Similarly, Kim et al implemented latent feature representation of breast lesions on DBT on a dataset independent from the one used in this study, and observed an AUC of 0.919 in characterizing breast masses, which agrees with the results of this study (38).

These promising preliminary results are encouraging and motivate future studies to evaluate the robustness of these findings in larger datasets, and to improve on these results through advances in techniques involved. While this study looked at data from multiple sources and views, we plan to merge these data sources to form a single impression on each subject. By incorporating all available data for a given lesion, we expect to see improvements in classification performance.

Furthermore, we acknowledge that the use of a key slice of the DBT volume is not necessarily optimal for this classification task. Before clinical implementation, efforts should be made to develop a system for optimized slice selection.

Alternatively, future methods could incorporate information from the full DBT volume. Incorporation of full volume information is expected to reduce bias introduced by selecting a single slice by involving all DBT image data available for the lesion of interest.

We plan to extend the scope of feature calculation algorithms beyond deep learning in order to compare and merge traditional handcrafted lesion features with those extracted by deep learning. Comparing a standard radiomics approach to the CNN-based approach taken in this study may explicate whether these algorithms extract redundant or complementary information. Understanding the relationship between these algorithms may be constructive in developing CADx systems for clinical use in breast imaging. While such an investigation would clearly be of value, this study omitted such a comparison as it focused instead on comparing value of breast image types, as opposed to comparing computer vision algorithm methodologies. As more images are collected at our institution, we plan to incorporate more sophisticated deep learning methods such as fine tuning and training from scratch. By continually improving computer-aided diagnosis of breast lesions, we hope to improve diagnostic accuracy and patient management for breast cancer patients.

ACKNOWLEDGMENTS

Supported, in part, by the NIBIB of the NIH under grant number T32 EB002103, the NCI of the NIH under grant number NIH QIN U01 195564. M.L.G. is a stockholder in R2 Technology/Hologic and a cofounder and shareholder in Quantitative Insights. M.L.G. and H.L receive royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi, and Toshiba. It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest that would reasonably appear to be directly and significantly affected by the research activities.

REFERENCES

1. Gur D, Abrams GS, Chough DM, et al. Digital breast tomosynthesis: observer performance study. Am J Roentgenol. Aug. 2009; 193:586–591.
2. Gennaro G, Toledano A, di Maggio C, et al. Digital breast tomosynthesis versus digital mammography: a clinical performance study. Eur Radiol. Jul. 2010; 20:1545–1553.
3. Wallis MG, Moa E, Zanca F, et al. Two-view and single-view tomosynthesis versus full-field digital mammography: high-resolution x-ray imaging observer study. Radiology. Mar. 2012; 262:788–796.
4. Haas BM, Kalra V, Geisel J, et al. Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. Radiology. Dec. 2013; 269:694–700.
5. Andersson I, Ikeda DM, Zackrisson S, et al. Breast tomosynthesis and digital mammography: a comparison of breast cancer visibility and BIRADS classification in a population of cancers with subtle mammographic findings. Eur Radiol. Dec. 2008; 18:2817–2825.
6. Reiser I, Nishikawa RM, Giger ML, et al. Computerized mass detection for digital breast tomosynthesis directly from the projection images. Med Phys. Feb. 2006; 33:482–491.
7. Mazurowski MA, Lo JY, Harrawood BP, Tourassi GD. Mutual information-based template matching scheme for detection of breast masses: From mammography to digital breast tomosynthesis.. J Biomed Inform 2011; 44(5):815–823
8. van Schie G, Wallis MG, Leifland K, et al. Mass detection in reconstructed digital breast tomosynthesis volumes with a computer-aided detection system trained on 2D mammograms. Med Phys. p. n/a-n/a, Apr. 2013; 40.
9. Giger ML, Karssemeijer N, Schnabel JA. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annu Rev Biomed Eng 2013; 15:327–357.
10. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. Advances in neural information processing systems, 25. Curran Associates, Inc; 2012. p. 1097–1105.
11. Bar Y, Diamant I, Wolf L, et al. Chest pathology detection using deep learning with non-medical training. In: 2015 IEEE 12th international symposium on biomedical imaging (ISBI); 2015. p. 294–297.
12. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. May 2016; 35:1285–1298.
13. Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition; 2009. p. 248–255.
14. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Trans Knowl Data Eng 2010; 22(10):1345–1359
15. Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J Med Imaging. Aug. 2016; 3:034501.
16. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med Phys. Oct. 2017; 44:5162–5171.
17. Samala RK, Chan HP, Hadjiski L, et al. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. Med Phys. Nov. 2016; 43:6654–6666.
18. Samala RK, Chan HP, Hadjiiski LM, et al. Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis. Med Imaging 2016: Comput Aided Diagn 2016; 9785:97850Y.
19. Samala RK, Chan HP, Hadjiiski LM, et al. Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. Phys Med Biol 2018; 63:095005.
20. K. Simonyan, A. Zisserman "Very deep convolutional networks for large-scale image recognition," ArXiv14091556 Cs, Sep. 2014.
21. L. Zheng, Y. Zhao, S. Want et al., "Good practice in CNN feature transfer," ArXiv160400133 Cs, Apr. 2016.
22. Draper NR. Applied regression analysis. 3rd ed. New York: Wiley, 1998.
23. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. Aug. 2005; 27:1226–1238.
24. Hua J, Xiong Z, Lowey J, et al. Optimal number of features as a function of sample size for various classification rules. Bioinform. Oxf Engl. Apr. 2005; 21:1509–1515.
25. Cortes C, Vapnik V. Support-vector networks. Mach Learn. Sep. 1995; 20:273–297.
26. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. Oct. 1978; 8:283–298.
27. Kittler J, Hatef M, Duin RPW, et al. On combining classifiers. IEEE Trans Pattern Anal Mach Intell. Mar. 1998; 20:226–239.
28. Gajdos C, Tartter P, Bleiweiss IJ, et al. Mammographic appearance of nonpalpable breast cancer reflects pathologic characteristics. Ann Surg. Feb. 2002; 235:246–251.
29. Friedman M. A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat 1940; 11:86–92
30. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. Sep. 1983; 148:839–843.
31. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979; 6:65–70.
32. Chan HP, Wei J, Sahiner B, et al. Computer-aided detection system for breast masses on digital tomosynthesis mammograms: preliminary experience. Radiology. Dec. 2005; 237:1075–1080.

33. Michell MJ, Iqbal A, Wasan RK, et al. A comparison of the accuracy of film-screen mammography, full-field digital mammography, and digital breast tomosynthesis. Clin Radiol. Oct. 2012; 67:976–981.
34. Morra L, Sacchetto D, Durando M, et al. Breast cancer: computer-aided detection with digital breast tomosynthesis. Radiology. May 2015; 277:56–63.
35. Park JM, Franken EA, Garg M, et al. Breast tomosynthesis: present considerations and future applications. Radiogr Rev Publ Radiol Soc N Am Inc. Oct. 2007; 27(Suppl 1):S231–S240.
36. Peppard HR, Nicholson BE, Rochman CM, et al. Digital breast tomosynthesis in the diagnostic setting: indications and clinical applications. RadioGraphics. May 2015; 35:975–990.
37. Baker JA, Lo JY. Breast tomosynthesis: state-of-the-art and review of the literature. Acad Radiol. Oct. 2011; 18:1298–1310.
38. Kim DH, Kim ST, Chang JM, et al. Latent feature representation with depth directional long-term recurrent learning for breast masses in digital breast tomosynthesis. Phys Med Biol 2017; 62:1009.